

A TAXONOMY OF INTELLIGENT INTRUSION DETECTION SYSTEMS FOR CYBERSECURITY USING LARGE LANGUAGE MODELS AND AGENT-BASED ARTIFICIAL INTELLIGENCE

Samson QUAYE*, Maurice DAWSON*, Enkel Hoxha**

*Center for Cybersecurity and Forensic Education (C²SAFE),
Illinois Institute of Technology, Chicago, USA

**Kolegji AAB, Kosovo
squaye@hawk.illinoistech.edu

Abstract: *Intrusion Detection Systems (IDS) play a critical role in modern cybersecurity, yet traditional and early artificial intelligence-based approaches remain limited by their reliance on predefined patterns, lack of contextual understanding, and high false positive rates. Recent advances in large language models (LLMs) and agent-based artificial intelligence introduce new capabilities that extend IDS beyond conventional detection toward reasoning, contextual awareness, and adaptive response. This work presents a taxonomy of intelligent intrusion detection systems for cybersecurity using LLMs and agent-based AI, organizing the design space across five dimensions: detection intelligence, functional role, system autonomy, data modality, and response capability. A unified architectural framework is also introduced, integrating LLM-based reasoning with agent-driven decision-making to support detection, explanation, and response. The paper highlights key advantages, including improved contextual correlation, explainability, and adaptive incident response, while addressing challenges related to reliability, security, operational constraints, and evaluation. It further outlines future directions toward autonomous, memory-augmented, and multi-agent IDS for complex environments.*

Keywords: Intrusion Detection Systems, Large Language Models, Artificial Intelligence, Agent-Based AI, Intelligent Cyber Defense, Cybersecurity

1. Introduction

Intrusion Detection Systems (IDS) have long served as a critical component of cybersecurity, providing mechanisms to monitor, analyze, and detect unauthorized or malicious activities within networked systems. Since their inception, IDS have evolved from rule-based and signature-driven approaches to more advanced anomaly detection techniques in response to the increasing sophistication of cyber threats. However, despite these advancements, modern cyber attacks continue to challenge traditional IDS due to their adaptive, multi-stage, and often low-signal nature. A fundamental limitation of conventional IDS lies in their reliance on

predefined signatures or statistical deviations, which restricts their ability to detect novel and evolving threats. As highlighted in prior work, a significant proportion of IDS alerts may be irrelevant or false positives, placing a considerable burden on security analysts and reducing operational efficiency [1]. While machine learning and deep learning approaches have improved detection accuracy and pattern recognition capabilities, they remain largely constrained by their dependence on feature engineering, limited contextual understanding, and lack of interpretability. Recent advances in Artificial Intelligence, particularly Large Language Models (LLMs) and agent-based systems, introduce

a paradigm shift in the design and functionality of IDS. Unlike traditional models, LLMs possess the ability to perform semantic reasoning over heterogeneous data sources, including network traffic, system logs, and threat intelligence reports. Furthermore, agent-based AI systems enable autonomous or semi-autonomous decision-making, allowing IDS to not only detect threats but also contextualize, explain, and respond to them in a dynamic manner. These capabilities extend IDS beyond static detection engines toward intelligent, adaptive cybersecurity systems.

Despite the growing interest in integrating advanced artificial intelligence techniques into intrusion detection systems, including machine learning and deep learning approaches, current research remains largely focused on enhancing existing IDS capabilities rather than redefining their underlying architectures [1]–[3]. Prior studies extensively explore improvements in detection accuracy, anomaly detection, and pattern recognition using learning-based models [3], [5], but these approaches remain constrained by limited contextual understanding and reliance on predefined features [4]. Furthermore, while neural network-based IDS have demonstrated significant advancements, existing literature does not provide a structured framework for understanding how emerging paradigms such as large language models and agent-based systems transform IDS design and functionality [2], [6]. As a result, there is a lack of comprehensive taxonomies that capture the integration of reasoning, contextual awareness, and autonomous decision-making in next-generation intrusion detection systems.

To address this gap, this paper proposes a comprehensive taxonomy of intelligent intrusion detection systems for cybersecurity based on large language models and agent-based artificial intelligence. Our research is guided by three primary questions: (1) How can intelligent IDS be systematically categorized to

capture the integration of LLMs and agent-based AI? (2) What architectural principles enable reasoning, memory, and adaptive decision-making in next-generation IDS? (3) What challenges must be addressed for practical deployment of such systems?

We employ a systematic literature analysis and architectural synthesis methodology. First, we review existing IDS literature spanning traditional, machine learning-based, and AI-enhanced approaches. Second, we synthesize this knowledge into a multi-dimensional taxonomy. Third, we propose a unified architectural framework integrating LLM-based reasoning with agent-driven decision-making. Finally, we identify implementation challenges and future research directions.

The contributions of this work are threefold. First, we introduce a unified taxonomy that categorizes intelligent IDS across multiple dimensions, including intelligence paradigm, functional role, autonomy level, data modality, and response capability. Second, we present a conceptual framework that formalizes the integration of reasoning, memory, and decision-making processes within IDS architectures. Third, we identify key challenges and outline future research directions toward the development of fully intelligent and autonomous intrusion detection systems.

This paper differs from prior IDS surveys by shifting the discussion from model performance to system intelligence. Rather than focusing only on classification accuracy, anomaly detection, or dataset comparison, this work examines how LLMs and agent-based AI reshape IDS architectures through semantic reasoning, contextual correlation, memory, tool use, and controlled autonomy. This perspective provides a foundation for understanding next-generation intrusion detection systems as intelligent cyber defense architectures rather than standalone detection models.

2. Background: From Traditional IDS to Intelligent IDS

Intrusion Detection Systems have evolved

significantly in response to increasingly sophisticated cyber threats. This section provides a concise overview of traditional IDS approaches, highlights their key limitations, and outlines the transition toward more intelligent systems driven by advances in artificial intelligence.

2.1 Traditional Intrusion Detection Systems

IDS have historically relied on two primary detection paradigms: signature-based detection and anomaly-based detection. Signature-based IDS identify malicious activities by matching observed patterns against a database of known attack signatures. This approach is highly effective for detecting previously identified threats and typically results in low false positive rates for known attack patterns. However, its effectiveness is inherently limited by its dependence on continuously updated signature databases, making it inadequate for detecting novel or zero-day attacks [2], [3].

In contrast, anomaly-based IDS establish a baseline of normal system or network behavior and identify deviations from this baseline as potential threats. These systems often employ statistical methods or machine learning techniques to model normal behavior and detect anomalies. While anomaly-based approaches are capable of identifying previously unseen attacks, they are prone to higher false positive rates, as benign variations in system behavior may be incorrectly classified as malicious [2], [10].

To mitigate the limitations of individual approaches, hybrid IDS combine signature-based and anomaly-based techniques, aiming to balance detection accuracy and generalization capability. Despite these improvements, traditional IDS architectures remain constrained in their ability to effectively address the complexity and scale of modern cyber threats [3].

2.2 Limitations of Traditional and Early AI-Based IDS

Despite significant advancements, both traditional and early AI-based IDS exhibit

several fundamental limitations. A key challenge is the reliance on predefined signatures or learned patterns, which restricts the ability of these systems to generalize to new and evolving attack vectors. As cyber threats become increasingly dynamic and sophisticated, this dependency limits the effectiveness of IDS in real-world environments [1], [3], [10]. Another major issue is the high rate of false positives, particularly in anomaly-based systems. The inability to accurately distinguish between benign and malicious deviations results in an overwhelming number of alerts, placing a substantial burden on security analysts and reducing operational efficiency [1], [4]. Additionally, machine learning-based IDS often depend on labeled datasets and feature engineering, which may not adequately capture the complexity of modern network behaviors [4], [9].

Furthermore, early AI-based IDS lack contextual awareness and interpretability. While deep learning models have demonstrated improved pattern recognition capabilities, they often operate as black-box systems, making it difficult for analysts to understand the reasoning behind detection decisions [5], [11], [12]. This limitation is particularly critical in cybersecurity, where explainability and trust are essential. Finally, existing IDS struggle to adapt in real time to emerging threats and to correlate information across heterogeneous data sources such as network traffic, system logs, and threat intelligence feeds. This limitation is exacerbated by the reliance on static datasets such as KDD Cup and UNSW-NB15, which may not fully capture the diversity and evolution of real-world cyber threats [7], [8].

2.3 The Shift to Intelligent Intrusion Detection Systems

The limitations of traditional and early AI-based IDS have driven the evolution toward more intelligent and adaptive cybersecurity systems. This progression can be understood as a transition from rule-based and signature-driven systems to machine

learning and deep learning models, and more recently, to systems enhanced by large language models and agent-based artificial intelligence.

Machine learning-based IDS introduced the ability to learn patterns from data, while deep learning approaches further improved the capacity to model complex and high-dimensional relationships in network traffic [11], [12]. However, these systems largely remain focused on classification and pattern recognition tasks, with limited capability for reasoning, contextual understanding, and autonomous decision-making [14].

Recent advances LLMs provide a new direction for IDS by enabling semantic analysis and reasoning over diverse and unstructured data sources [16]. LLMs can process and correlate information across logs, alerts, and threat intelligence, offering deeper contextual understanding of potential security incidents. In parallel, agent-based artificial intelligence introduces the concept of autonomous or semi-autonomous systems capable of decision-making, planning, and interaction with external tools and environments.

Together, these developments mark a shift toward intelligent intrusion detection systems that extend beyond static detection mechanisms. Such systems are increasingly capable of integrating reasoning, memory, and adaptive response, laying the foundation for next-generation IDS architectures that are more resilient, context-aware, and responsive to evolving cyber threats.

3. Taxonomy of Intelligent Intrusion Detection Systems

The rapid evolution of artificial intelligence has led to the emergence of intrusion detection systems that extend beyond traditional pattern recognition toward reasoning, contextual understanding, and adaptive response. While prior studies have explored the application of machine learning and deep learning techniques in IDS [1], [3], [10], existing literature largely presents these advancements in a

fragmented manner without a structured framework to systematically characterize their capabilities. To provide clarity, intelligent intrusion detection systems can be organized along five key dimensions: detection intelligence, functional role, system autonomy, data modality, and response capability. The proposed taxonomy is illustrated in Fig. 1, highlighting the multi-dimensional structure of intelligent intrusion detection systems across detection intelligence, functional roles, system autonomy, data modalities, and response capabilities.

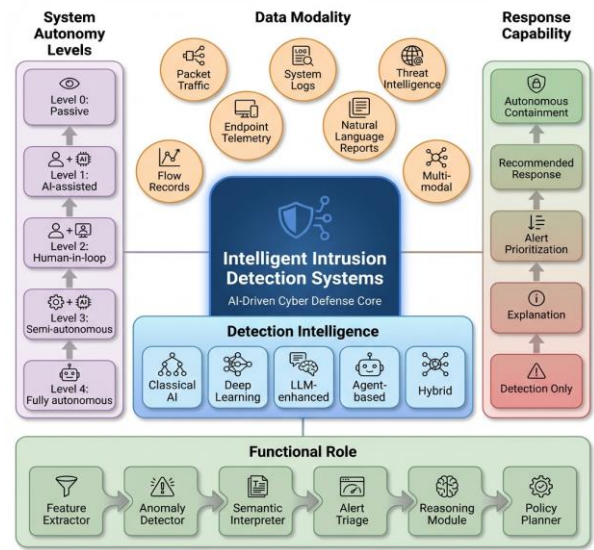


Fig. 1. Taxonomy of intelligent intrusion detection systems across detection intelligence, functional roles, autonomy, data, and response. (Original figure developed by authors based on systematic analysis of IDS literature)

3.1 Taxonomy by Detection Intelligence

The first dimension categorizes IDS based on the underlying intelligence paradigm that drives detection and analysis. Classical AI-assisted IDS rely on machine learning techniques such as decision trees, clustering, and support vector machines to identify patterns in network traffic [10], [3]. Deep learning-based IDS extends this capability by leveraging neural networks to model complex and high-dimensional relationships, significantly improving detection performance in large-scale environments [11], [12], [14]. More recently, LLM-enhanced IDS introduced

semantic reasoning capabilities, enabling systems to analyze unstructured and heterogeneous data such as logs, alerts, and threat intelligence reports. This represents a shift from purely statistical detection toward contextual understanding, as also suggested by recent explorations of advanced AI integration in IDS [1], [16]. Agent-based IDS further extends this paradigm by enabling autonomous or semi-autonomous decision-making, where intelligent agents can plan, act, and interact with external tools to support detection and response processes. Hybrid intelligent IDS combine multiple paradigms, integrating machine learning, deep learning, and LLM-based reasoning to achieve improved adaptability and robustness.

3.2 Taxonomy by Functional Role of the Model

Beyond detection capability, intelligent IDS can be classified based on the functional roles performed by AI models within the system. At a foundational level, models act as feature extractors, transforming raw network or system data into structured representations suitable for analysis, a concept widely used in machine learning-based IDS [10], [11]. At the detection level, models function as anomaly detectors or classifiers, identifying deviations from normal behavior or known attack patterns [3], [12]. More advanced systems incorporate semantic interpretation, where models analyze contextual information across multiple data sources. In operational environments, AI models may serve as alert triage engines, addressing the well-documented issue of high alert volumes and false positives in IDS [1], [4]. Additionally, reasoning and explanation modules enable systems to provide interpretable insights into detected threats, addressing the limitations of black-box deep learning models [5], [6]. At the highest level, models act as policy and response planners, supporting decision-making processes related to threat mitigation and system defense.

3.3 Taxonomy by System Autonomy

A key differentiator of intelligent IDS lies in the degree of autonomy they exhibit. Traditional IDS are largely passive systems that generate alerts without adaptive intelligence [2], [3]. The introduction of machine learning and deep learning enabled AI-assisted IDS, where models support detection but require human interpretation and intervention [10], [14]. As systems evolve, analyst-in-the-loop IDS integrate human expertise with AI-driven insights, improving detection accuracy and decision-making. Semi-autonomous IDS further extend this capability by automating limited response actions such as alert prioritization and recommendation generation. Finally, controlled agent-based IDS represent an emerging paradigm, where systems can independently detect, analyze, and respond to threats with minimal human intervention. This progression aligns with broader trends in AI toward autonomous decision-making systems, while also introducing new challenges related to trust, safety, and control.

3.4 Taxonomy by Data Modality

Intelligent IDS must operate across diverse and heterogeneous data sources. Traditional IDS primarily rely on packet-level traffic and flow records for detection [2], [9]. Modern systems expand this scope to include system logs, endpoint telemetry, and host-based activity data, reflecting the increasing complexity of cyber environments [3], [14]. In addition, the integration of threat intelligence feeds and external data sources enables more comprehensive situational awareness. Recent datasets such as KDD Cup and UNSW-NB15 have been widely used to train and evaluate IDS models, although they may not fully capture real-world attack diversity [7], [8], [9]. With the emergence of LLM-based approaches, IDS can now incorporate natural language reports and unstructured data, enabling multi-modal analysis. Multi-modal IDS represent the most advanced category, combining heterogeneous data sources to detect complex, multi-stage attacks.

3.5 Taxonomy by Response Capability

The final dimension considers the response capabilities of intelligent IDS, reflecting their role within the broader cybersecurity workflow. Traditional IDS primarily function as detection-only systems, generating alerts without further analysis [2], [3]. More advanced systems provide detection with explanation, offering insights into the nature and cause of detected threats, addressing the interpretability challenges of AI-based IDS [5], [6]. Further progression includes detection with alert prioritization, which helps mitigate the high volume of alerts and reduces analyst workload [1], [4]. Detection with recommended response extends this capability by suggesting mitigation strategies based on observed threat patterns. At the highest level, intelligent IDS support detection with autonomous containment, where systems can initiate defensive actions such as blocking malicious traffic or isolating compromised systems. This capability reflects the broader evolution toward adaptive and self-defending cybersecurity systems.

4. Architecture of LLM and Agent-Based Intrusion Detection Systems

The transition toward intelligent intrusion detection systems requires architectures that integrate data-driven detection with contextual reasoning and adaptive response. Recent advances in large language models and autonomous AI systems have demonstrated strong capabilities in reasoning, planning, and multi-step decision-making across complex tasks [16], [17], [18]. Building upon the taxonomy presented in Section 3, this section outlines a unified architectural perspective that combines these paradigms to enable more effective and adaptive cybersecurity operations.

4.1 LLM-Enhanced IDS Pipeline

Modern intrusion detection systems increasingly incorporate large language models to enhance contextual understanding and reasoning over

heterogeneous data sources. LLMs have shown strong performance in processing unstructured data, including logs, reports, and security alerts, enabling semantic interpretation beyond traditional feature-based approaches [16],[19]. A typical LLM-enhanced IDS pipeline begins with data ingestion, where network traffic, system logs, and threat intelligence feeds are collected and aggregated. This is followed by a preprocessing stage, where raw data is normalized, filtered, and transformed into structured or semi-structured representations, consistent with established IDS data processing practices [10], [14]. Next, the processed data is converted into embeddings or structured representations, enabling efficient semantic analysis. A retrieval and context-building module augments the input with relevant historical data, aligning with retrieval-augmented generation techniques that enhance LLM performance in knowledge-intensive tasks [19], [20]. The LLM subsequently performs reasoning and analysis, generating insights such as anomaly interpretation, attack classification, or threat explanation. The outputs of this stage are passed to an alert classification and explanation module, where detected events are categorized and accompanied by interpretable justifications. Finally, a response recommendation component suggests potential mitigation strategies, supporting decision-making processes in cybersecurity operations.

A policy evaluation module ensures that selected actions comply with predefined security policies and operational constraints. The pipeline concludes with a mitigation and containment module, where the system executes or recommends actions such as blocking malicious traffic, isolating compromised components, or escalating incidents to human analysts.

4.2 Agent-Based IDS Pipeline

While LLM-enhanced pipelines improve reasoning, agent-based architectures extend IDS capabilities toward autonomy and adaptive response. Recent work on

autonomous agents highlights their ability to perform multi-step reasoning, tool usage, and environment interaction, making them well-suited for complex cybersecurity tasks [17], [18], [21]. The pipeline begins with sensor inputs, including network telemetry, logs, and external intelligence sources. These inputs are used to construct a state representation of the system, capturing the current security posture. A memory and context store maintains historical observations, prior incidents, and learned patterns, enabling temporal reasoning and long-term adaptation. An agent planner determines appropriate actions based on the current state and objectives. This is supported by a tool usage layer, where the agent can invoke external tools such as intrusion detection engines, forensic analyzers, or threat intelligence services. Such tool-augmented reasoning has been shown to significantly enhance decision-making capabilities in complex environments [17], [21]. This design is consistent with recent guided AI approaches for intrusion detection and automated response in critical infrastructure environments, where LLM-based IDS components are explored for anomaly detection, threat analysis, and response support [23].

4.3 Human-in-the-Loop Security Operations

Despite advances in automation, human expertise remains a critical component of intelligent intrusion detection systems. Human analysts are essential for validating high-risk actions, particularly in scenarios where incorrect decisions may lead to operational disruption or security risks. This aligns with established findings that human oversight improves the reliability and trustworthiness of AI-driven systems [4], [22]. Additionally, analysts provide feedback for continual learning, enabling systems to adapt to evolving threats. Human-in-the-loop approaches are particularly important in mitigating challenges associated with LLMs and agent-based systems, including

hallucination, bias, and model drift [16], [22]. By integrating human oversight into the decision-making loop, IDS architectures can achieve a balance between automation and accountability.

4.4 Conceptual Formulation

The operation of an intelligent intrusion detection system can be viewed as a function that integrates observed telemetry, contextual information, and historical knowledge to produce detection outcomes and threat assessments. In this formulation, the system processes incoming data such as network traffic and system logs, augments it with contextual threat intelligence, and leverages stored memory from prior incidents to improve decision-making. Similarly, in agent-based IDS, the response process can be interpreted as a policy-driven decision mechanism in which actions are selected based on the current system state and historical context. This enables the system to adapt its behavior dynamically, supporting tasks including alert prioritization, response recommendation, and automated mitigation. This formulation summarizes the intelligent IDS process as a controlled pipeline that links observation, reasoning, memory, and policy-constrained response.

5. Advantages of Intelligent Intrusion Detection Systems over Earlier AI-Based IDS

Intelligent intrusion detection systems extend the capabilities of earlier AI-based approaches by incorporating semantic reasoning, contextual awareness, and adaptive decision-making. While traditional machine learning and deep learning models primarily focus on pattern recognition, they often lack the ability to interpret complex security events in context. In contrast, LLM- and agent-based IDS enable semantic reasoning over alerts and logs, allowing systems to understand relationships between events rather than treating them as isolated observations. A key advantage lies in the ability to perform contextual correlation across heterogeneous

data sources, including network traffic, system logs, and threat intelligence. This enables more accurate detection of sophisticated and multi-stage attacks that span multiple domains. Furthermore, intelligent IDS improve explainability by providing human-readable interpretations of detected threats, which enhances analyst trust and supports more informed decision-making.

Another important benefit is adaptive incident triage, where alerts can be prioritized dynamically based on severity, context, and historical patterns. This reduces the cognitive burden on security analysts and improves response efficiency. In addition, agent-based architectures introduce support for semi-autonomous and autonomous response mechanisms, enabling systems to recommend or execute mitigation actions in real time while adhering to defined policies. Finally, intelligent IDS demonstrate improved capability in handling zero-day and evolving threats, as they are not strictly limited to predefined signatures or static learned patterns. By integrating reasoning, memory, and contextual understanding, these systems provide a more flexible and robust defense against modern cyber threats.

6. Challenges, Risks, and Open Problems

Despite the advantages of intelligent intrusion detection systems, the integration of large language models and agent-based architectures introduces several challenges that must be carefully addressed. These limitations span reliability, security, operational constraints, and evaluation, and they highlight the gap between current capabilities and fully trustworthy deployment in real-world environments.

6.1 Reliability Risks

LLM- and agent-based IDS are susceptible to reliability issues arising from the inherent behavior of generative models. These include hallucinations, where the system may produce plausible but incorrect interpretations of security events, and

inconsistent reasoning across similar inputs. Additionally, model outputs can be sensitive to input formulation, leading to variability in detection and explanation. In agent-based pipelines, such issues may propagate across multiple stages, amplifying errors and potentially resulting in incorrect or unsafe decisions.

6.2 Security Risks

The introduction of LLMs and agents expands the attack surface of IDS. Adversarial prompts may manipulate model behavior, causing incorrect classifications or misleading explanations. Data poisoning attacks can compromise training or contextual data sources, degrading system performance. Furthermore, risks such as model extraction and unauthorized access to internal reasoning processes raise concerns about intellectual property and system integrity. In agent-based systems, unsafe or unintended action triggering presents an additional challenge, particularly when systems are granted partial autonomy.

6.3 Operational Constraints

Practical deployment of intelligent IDS is constrained by computational and system-level limitations. LLM inference introduces latency that may conflict with real-time detection requirements, especially in high-throughput network environments. In addition, the cost of running large-scale models can be significant, particularly when continuous monitoring is required. Memory and context management also pose challenges, as maintaining long-term historical information can lead to scalability issues in large and dynamic systems.

6.4 Privacy and Governance

The use of sensitive security data raises important privacy and governance concerns. Logs and telemetry may contain confidential or regulated information, requiring careful handling and compliance with data protection standards. The integration of LLMs further complicates this, as external model APIs or shared infrastructures may introduce additional exposure risks. Moreover, explainability

and accountability become critical when systems generate automated decisions, particularly in regulated environments where transparency and traceability are required.

6.5 Evaluation Problems

Evaluating intelligent IDS remains an open problem. Existing benchmarks and datasets are primarily designed for traditional machine learning approaches and may not capture the complexity of LLM reasoning or agent-based behavior. There is often a mismatch between academic datasets and real-world security operations center data, limiting the validity of experimental results. Recent work on AI-powered cybersecurity models for IoT devices further highlights the importance of evaluating models across diverse datasets and deployment conditions rather than relying on isolated or outdated benchmarks [24]. Furthermore, current evaluation metrics focus on detection accuracy but do not adequately assess reasoning quality, explanation usefulness, or the effectiveness of automated response actions.

7. Future Directions for Next-Generation Intrusion Detection Systems

The evolution of intelligent intrusion detection systems is expected to move toward more adaptive, autonomous, and context-aware capabilities. One important direction is the development of self-improving IDS that continuously learn from analyst feedback, incident outcomes, and post-incident analysis. Such systems can refine detection strategies over time, improving accuracy and reducing false positives in dynamic threat environments. Another promising direction lies in multi-agent security ecosystems, where specialized agents collaborate across sensing, reasoning, validation, and response orchestration. This distributed approach enables more scalable and modular security architectures, allowing different components to operate concurrently while sharing contextual knowledge. In parallel, memory-augmented IDS are expected to

play a significant role by maintaining persistent knowledge of past attacks, adversary behavior, and long-term system states, enabling improved detection of recurring and multi-stage threats. Generative AI also introduces new opportunities for threat-informed simulation, where synthetic attack scenarios can be generated to test and evaluate IDS performance under rare or emerging threat conditions. This can support continuous system validation and improve robustness against previously unseen attack patterns. At the same time, ensuring trust and safety in autonomous systems remains critical. Future IDS will likely incorporate policy-constrained decision-making, human approval layers, and verifiable safeguards to ensure that automated responses remain controlled and accountable. Finally, the growing complexity of modern infrastructures suggests a need for cross-domain intelligent IDS capable of operating across enterprise networks, cloud environments, Internet of Things ecosystems, and cyber-physical systems. Such unified systems would provide a holistic view of security posture, enabling coordinated detection and response across diverse and interconnected environments.

8. Conclusion

The evolution of intrusion detection systems reflects a shift from traditional signature-based and machine learning approaches toward more intelligent, context-aware, and adaptive cybersecurity frameworks. The integration of large language models and agent-based intelligence introduces new capabilities in semantic reasoning, contextual understanding, and automated decision-making, enabling IDS to address increasingly complex and dynamic threat environments. At the same time, these advancements introduce challenges related to reliability, security, operational constraints, and governance, which must be carefully addressed to ensure safe and effective deployment. Managing these

challenges is essential for transitioning intelligent IDS from conceptual designs to practical systems. The future of intrusion detection systems lies not merely in improved classification accuracy, but in the integration of reasoning, memory, explainability, and controlled autonomy into a unified cybersecurity architecture. Specifically, key research directions include developing evaluation frameworks that assess reasoning quality beyond

classification metrics, establishing trust and safety mechanisms for autonomous IDS with formal verification, advancing memory-augmented architectures that balance long-term knowledge with computational constraints, designing multi-agent security ecosystems with verifiable coordination, and creating cross-domain intelligent IDS for enterprise, cloud, and cyber-physical systems.

References

- [1] Markevych, M., Dawson, M. A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (AI). In: International Conference Knowledge-Based Organization. Knowledge-Based Organization Conference Proceedings. Romania: Land Forces Academy Publishing House; 2023. pp. 30–37.
- [2] Liao, H. J., Lin, C. H. R., Lin, Y. C., Tung, K. Y. Intrusion detection system: A comprehensive review. In: Journal of Network and Computer Applications. Amsterdam: Elsevier; 2013. pp. 16–24.
- [3] Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J. Survey of intrusion detection systems: techniques, datasets and challenges. In: Cybersecurity. London: Springer; 2019. pp. 1–22.
- [4] Sommer, R., Paxson, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In: IEEE Symposium on Security and Privacy. California: IEEE; 2010. pp. 305–316.
- [5] Kim, A., Park, M., Lee, D. H. AI-IDS: Application of deep learning to real-time web intrusion detection. In: IEEE Access. New York: IEEE; 2020. pp. 70245–70261.
- [6] Drewek-Ossowicka, A., Pietrolaj, M., Rumiński, J. A survey of neural networks usage for intrusion detection systems. In: Journal of Ambient Intelligence and Humanized Computing. Berlin: Springer; 2020. pp. 497–514.
- [7] Tavallae, M., Bagheri, E., Lu, W., Ghorbani, A. A. A detailed analysis of the KDD CUP 99 data set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications. USA: IEEE; 2009. pp. 1–6.
- [8] Moustafa, N., Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems. In: Military Communications and Information Systems Conference. Australia: IEEE; 2015. pp. 1–6.
- [9] Ring, M., Wunderlich, S., Grödl, D., Landes, D., Hotho, A. A survey of network-based intrusion detection data sets. *Computers & Security*. Oxford: Elsevier; 2019. pp. 147–167.
- [10] Buczak, A. L., Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*. New York: IEEE; 2016. pp. 1153–1176.
- [11] Shone, N., Ngoc, T. N., Phai, V. D., Shi, Q. A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*. New York: IEEE; 2018. pp. 41–50.
- [12] Yin, C., Zhu, Y., Fei, J., He, X. A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access*. New York: IEEE; 2017. pp. 21954–21961.

- [13] Vinayakumar, R., Soman, K. P., Poornachandran, P. Applying deep learning approaches for network traffic prediction. In: International Conference on Advances in Computing, Communications and Informatics. India: IEEE; 2017. pp. 2353–2358.
- [14] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., Janicke, H. Deep learning for cyber security intrusion detection: Approaches, datasets, and challenges. *Journal of Information Security and Applications*. Amsterdam: Elsevier; 2020. pp. 102419.
- [15] Goodfellow, I., Shlens, J., Szegedy, C. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations. 2015.
- [16] OpenAI. GPT-4 Technical Report. arXiv preprint arXiv: 2303. 08774; 2023.
- [17] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y. ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR). 2023.
- [18] Xi, Z., Chen, W., Guo, X., et al. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864; 2023.
- [19] Lewis, P., Perez, E., Piktus, A., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS). 2020.
- [20] Gao, L., Ma, X., Lin, J., Callan, J. Precise zero-shot dense retrieval without relevance labels. In: Annual Meeting of the Association for Computational Linguistics (ACL). 2022.
- [21] Schick, T., Dwivedi-Yu, J., Dessì, R., et al. Toolformer: Language models can teach themselves to use tools. In: Advances in Neural Information Processing Systems (NeurIPS). 2023.
- [22] Amershi, S., Weld, D., Vorvoreanu, M., et al. Guidelines for human-AI interaction. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). New York: ACM; 2019.
- [23] Zziwa, I., Pawar, H., Nartey, C., Dawson, M. GUARD: A Guided AI System for Intrusion Detection and Automated Response in Critical Infrastructure Environments. In: International Conference Knowledge-Based Organization. 2025; 31(1). DOI: 10.2478/kbo-2025-0028.
- [24] Quaye, S., Khan, A. H., Tapre, K., Dawson, M. AI-Powered Cybersecurity Models for Training and Testing IoT Devices. *Applied Sciences*. 2025; 15(24):13073. DOI: 10.3390/app152413073.